Trabalho de Regressão II

Mariana Costa freitas

2024-06-06

Introdução

O diabetes é uma condição crônica que afeta milhões de pessoas em todo o mundo, e seu diagnóstico precoce é fundamental para a prevenção de complicações graves. Nesse contexto, o Instituto Nacional de Diabetes e Doenças Digestivas e Renais realizou um estudo com mulheres Pimas adultas residentes na região de Phoenix, Arizona, com o objetivo de investigar fatores que possam influenciar a concentração de glicose no sangue, medida através de um teste oral de tolerância à glicose.

Este relatório apresenta uma análise de regressão para a variável Glucose, a fim de identificar os principais fatores associados a suas variações. As variáveis utilizadas no estudo incluem características fisiológicas e dados demográficos das participantes:

O objetivo principal deste estudo é construir um modelo de regressão linear para a variável Glucose e verificar a adequação desse modelo através da análise dos pressupostos fundamentais da regressão linear, tais como linearidade, independência dos resíduos, homocedasticidade e normalidade dos resíduos.

Ao final, espera-se identificar as variáveis que mais influenciam a concentração de glicose no sangue e fornecer insights valiosos para estratégias de prevenção e tratamento do diabetes entre as mulheres Pimas.

Análise dos dados

Nessa análise estamos utilizando os dados disponibilizados pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais, obtidos por meio de uma pesquisa feita em índias Pimas adultas que vivem perto de Phoenix-Arizona. As variáveis abordadas no banco de dados são:

- pregnancies: número de gestações
- glucose: concentração de glicose em teste oral de tolerância à glicose
- blood_pressure: pressão arterial diastólica (mm Hg)
- skin_thickness: espessura cutânea triciptal (mm)
- insulin: 2 horas de insulina no soro (mu U/ml)
- bmi: índice de massa corporal (IMC)
- diabetes_pedigree_function: função que mede as chances de ter diabetes baseada no histórico familiar
- age: idade em anos
- outcome: resultado do teste para diabetes, que pode ser saudável (outcome=0) ou diabético (outcome=1)

Análise descritiva

Para melhor compreensão dos dados que estamos utilizando, vamos obter algumas de suas medidas descritivas:

	N	Mean	SD	Min	Q1	Median	Q3	Max
pregnancies	768	3.85	3.37	0.00	1.00	3.00	6.00	17.00
glucose	768	120.89	31.97	0.00	99.00	117.00	140.50	199.00
blood_pressure	768	69.11	19.36	0.00	62.00	72.00	80.00	122.00
skin_thickness	768	20.54	15.95	0.00	0.00	23.00	32.00	99.00
insulin	768	79.80	115.24	0.00	0.00	30.50	127.50	846.00
bmi	768	31.99	7.88	0.00	27.30	32.00	36.60	67.10
diabetes_pedigree_function	768	0.47	0.33	0.08	0.24	0.37	0.63	2.42
age	768	33.24	11.76	21.00	24.00	29.00	41.00	81.00
outcome	768	0.35	0.48	0.00	0.00	0.00	1.00	1.00

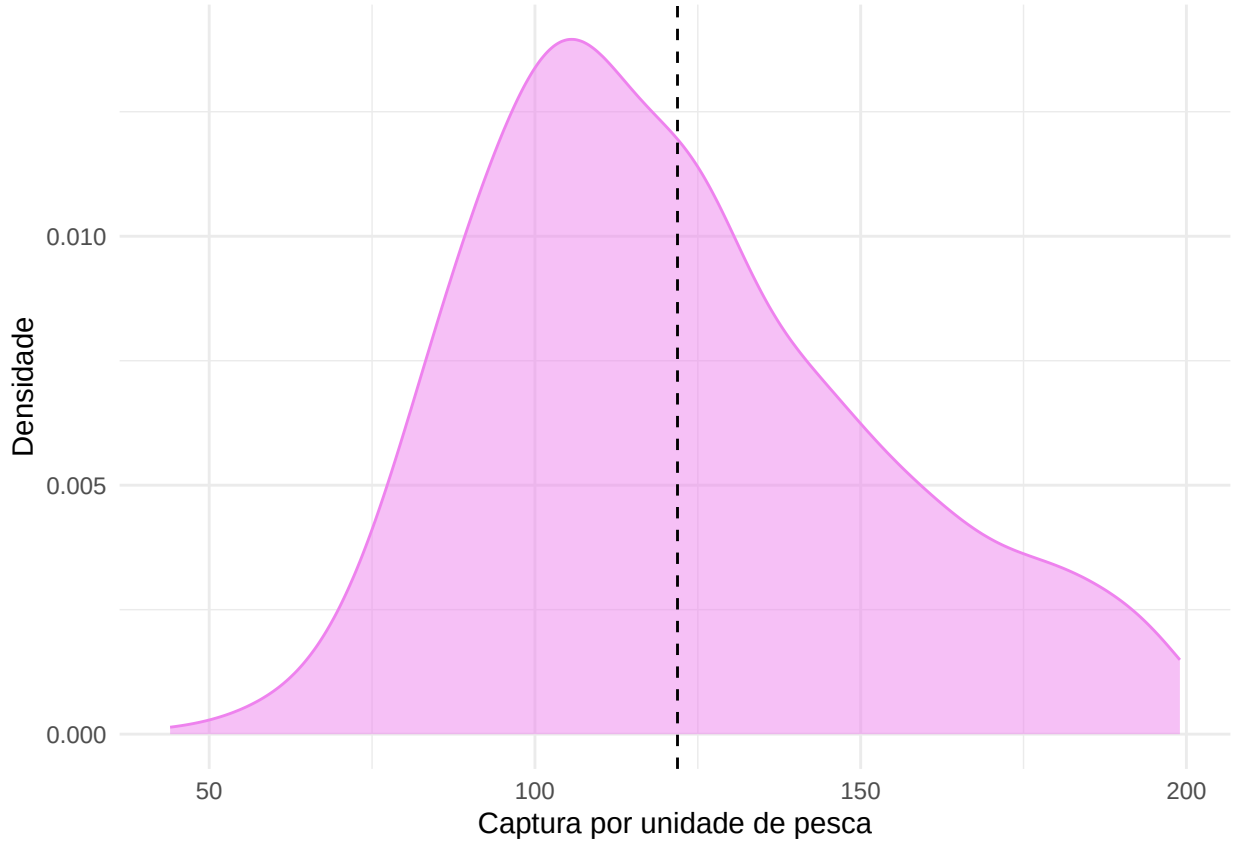
Essas variáveis apresentam muitos dados faltantes que podem prejudicar nossas análises e posteriormente também o desenvolvimento do modelo de regressão. Assim, é necessário remover od dados faltantes, e remover aquelas variáveis que apresentam alta quantidade desses dados.

Variável	Frequência de dados faltantes
glucose	5
blood_pressure	35
skin_thickness	227
insulin	374
bmi	11
age	0
pregnancies	0
outcome	0
diabetes_pedigree_function	0

Dessa forma, vamos excluir as variáveis `insulin` e `skin_thicknesse` remover os dados faltantes de `glucose`, `blood_pressure` e `bmi`. O restante das variáveis não apresentam dados faltantes, então não vamos alterá-las. Abaixo podemos observar as novas medidas descritivas após o tratamento desses dados.

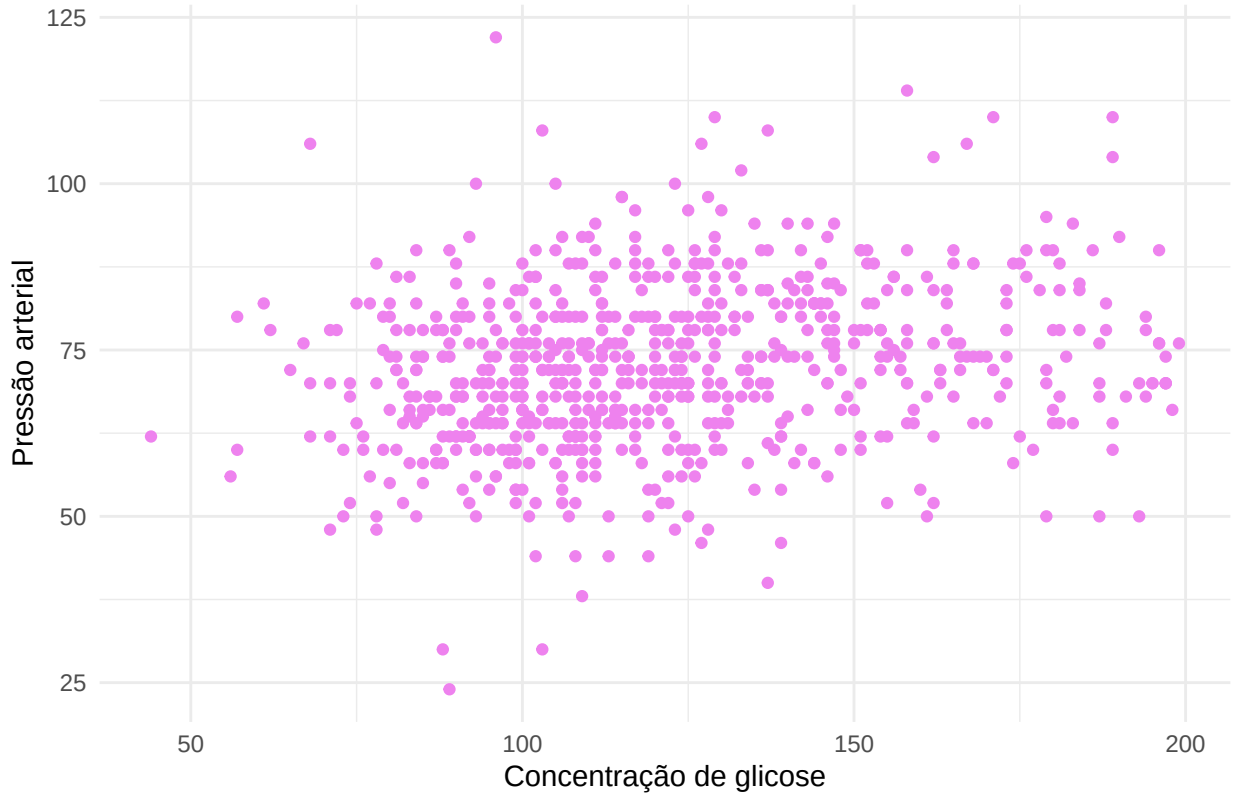
	N	Mean	SD	Min	Q1	Median	Q3	Max
pregnancies	724	3.87	3.36	0.00	1.00	3.00	6.00	17.00
glucose	724	121.88	30.75	44.00	99.50	117.00	142.00	199.00
blood_pressure	724	72.40	12.38	24.00	64.00	72.00	80.00	122.00
bmi	724	32.47	6.89	18.20	27.50	32.40	36.60	67.10
diabetes_pedigree_function	724	0.47	0.33	0.08	0.24	0.38	0.63	2.42
age	724	33.35	11.77	21.00	24.00	29.00	41.00	81.00
outcome	724	0.34	0.48	0.00	0.00	0.00	1.00	1.00

Um importante passo na análise descritiva é verificar o comportamento da variável `glucose`, que será a variável descrita pelo modelo de regressão. Abaixo, construímos um histograma para melhor entender a sua distribuição.

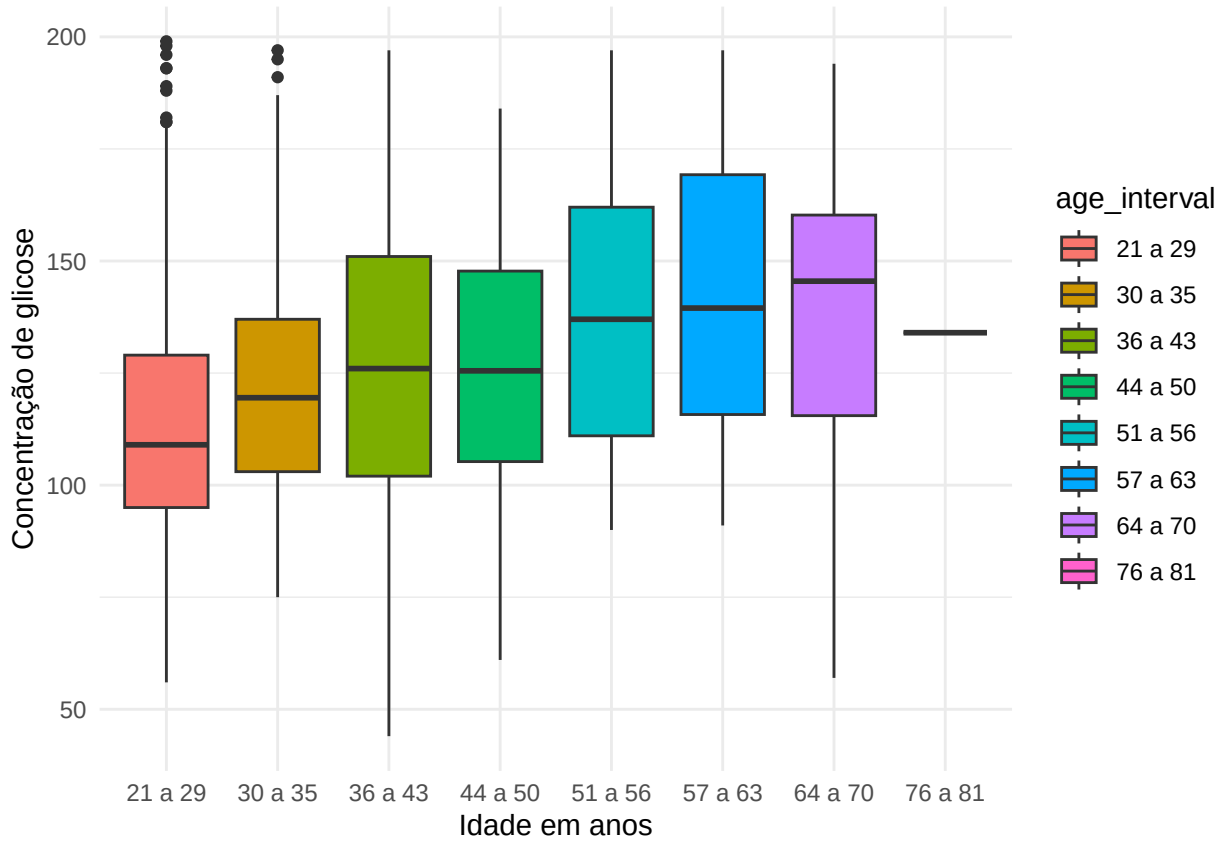


Aqui percebemos que seu gráfico de densidade apresenta leve assimetria a direita, então futuramente talvez seja necessário realizar uma transformação de Box-Cox.

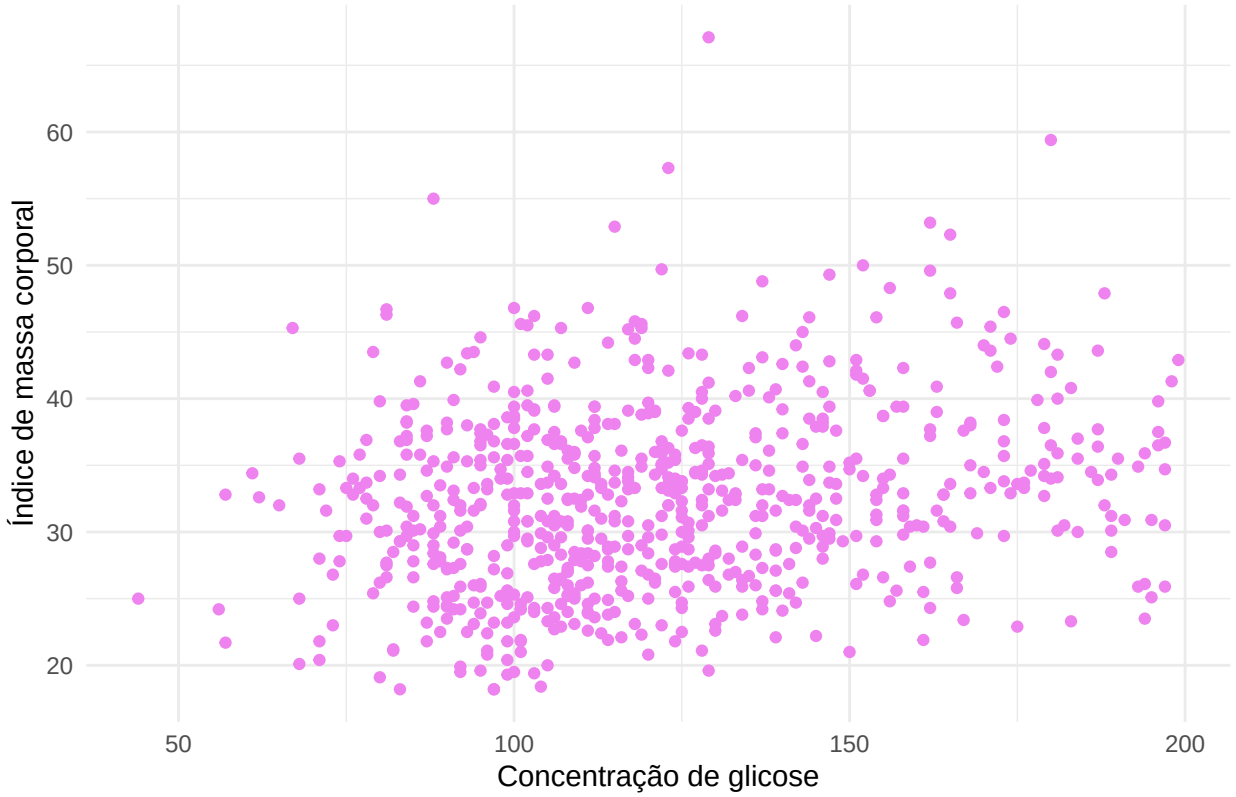
Também vamos observar o comportamento de `glucose` associada a outras variáveis abaixo.



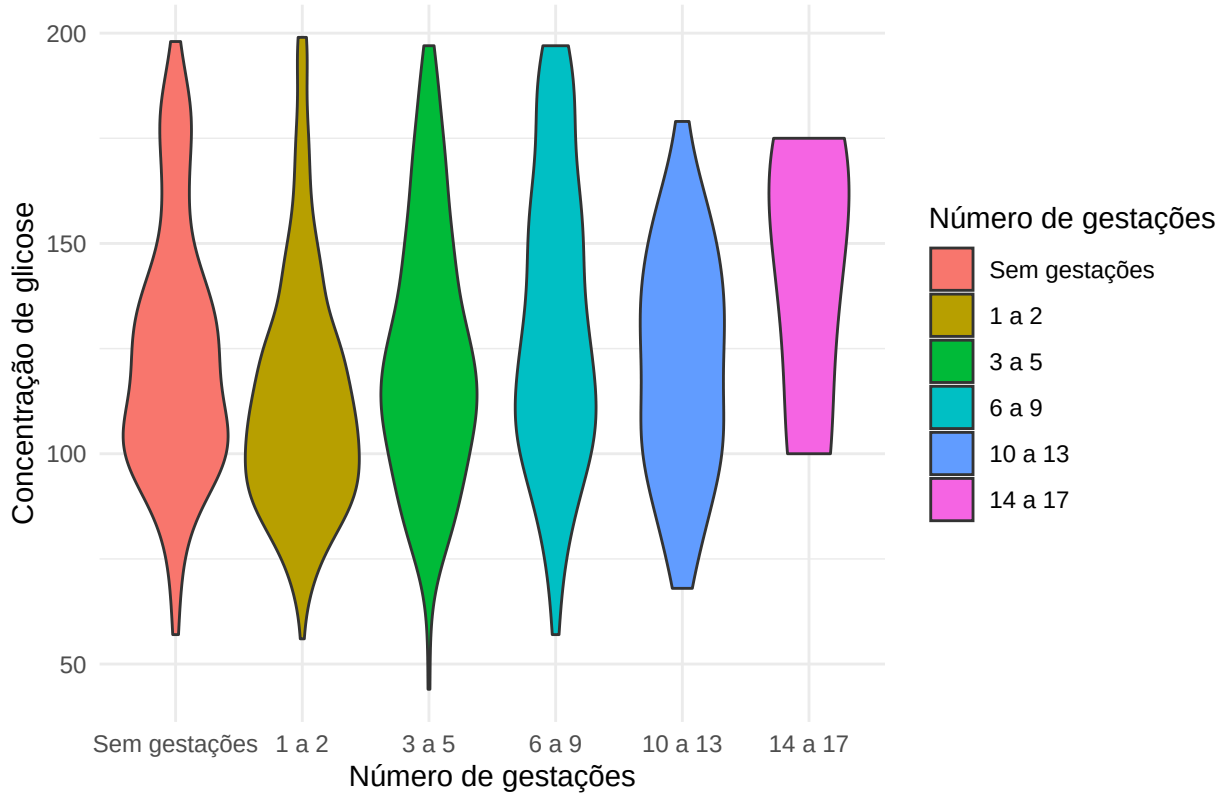
É possível perceber que o comportamento da variável *glucose* não é influenciado pela variável *blood_pressure* e vice-versa, já que não observamos nenhum padrão na distribuição de uma conforme aumento ou diminuição da outra.



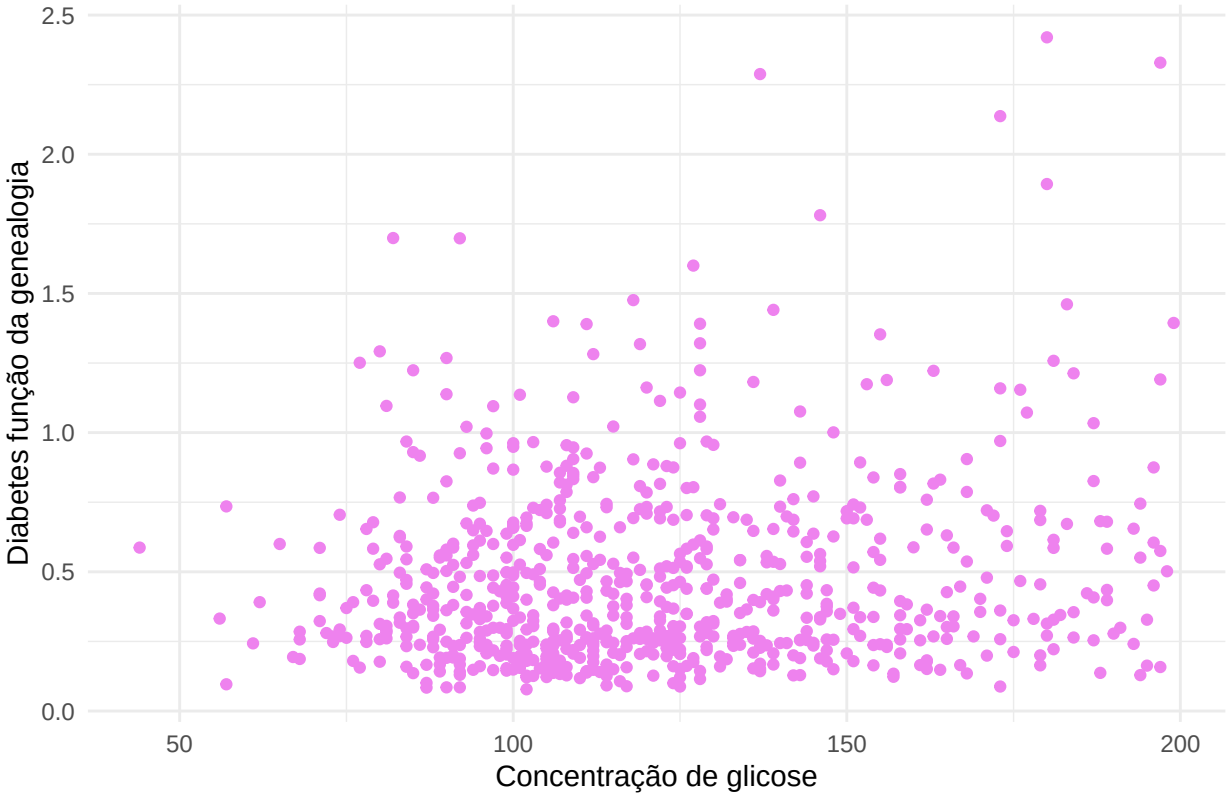
No boxplot acima, discretizamos a variável que representa idade, por meio de um agrupamento em intervalos. Observamos que, a medida que a idade aumenta, a média da concentração de glicose também aumenta gradualmente, além de apresentar variabilidade levemente maior.



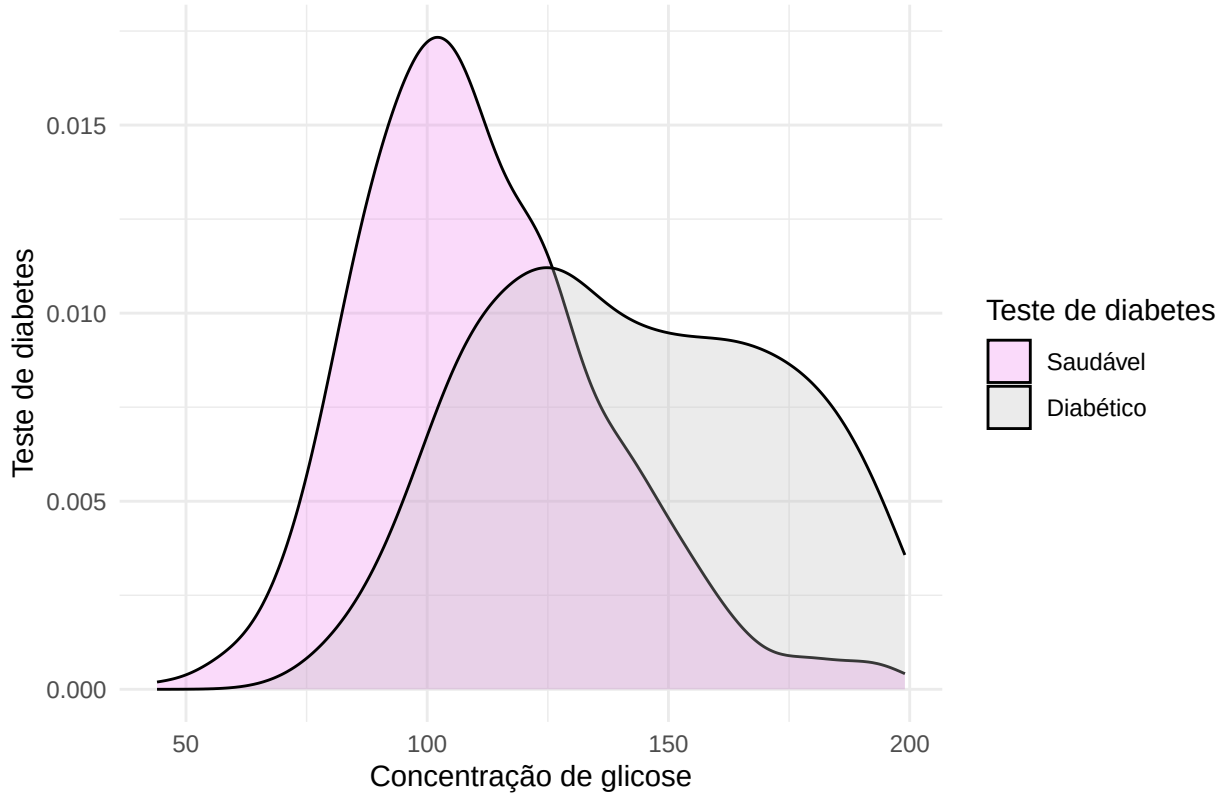
Na representação acima, podemos observar que não há uma tendência a aumento ou diminuição de uma das variáveis enquanto a outra aumenta, o IMC se mantém distribuído da mesma forma conforme a variação da concentração de glicose.



Nesse gráfico, as idades gestacionais (em semanas) foram agrupadas em intervalos. A partir do gráfico de violino, é possível notar que nos primeiros grupos a glicose se concentra em níveis mais baixos, já a partir das semanas 6 a 9 o nível de glicose começa a se distribuir mais uniformemente, enquanto no último grupo se concentra em níveis de glicose mais altos.



Aqui podemos notar que aparentemente não há uma linha reta ou curva que descreva a relação entre as variáveis. Os pontos não seguem uma inclinação positiva (indicando correlação positiva) ou negativa (indicando correlação negativa).



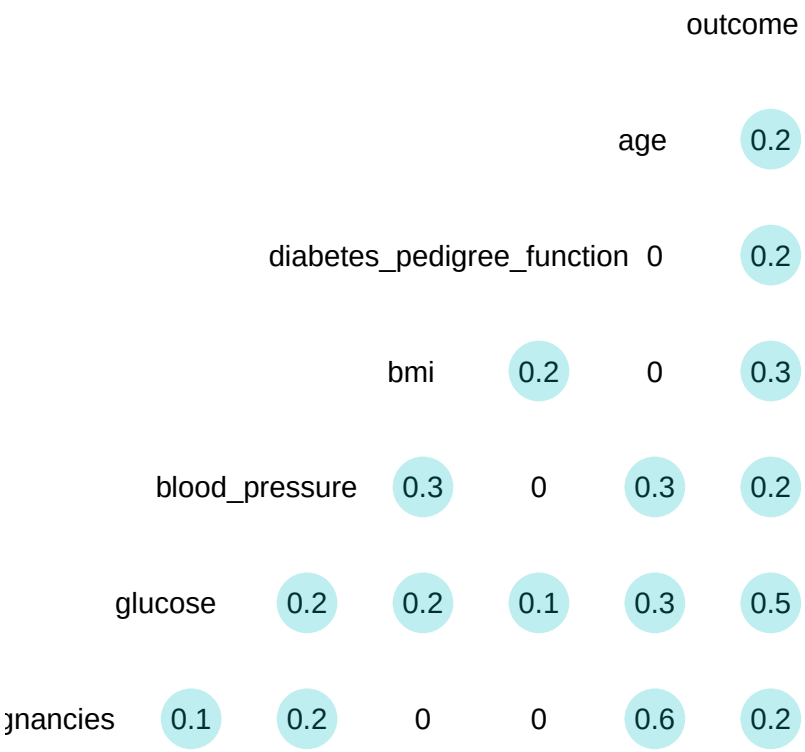
Acima temos as curvas de densidade para o nível de glicose, segregado em saudável e diabético. A partir do gráfico de densidade podemos observar que há grande diferença nos níveis de glicose entre os dois grupos, tendo os diabéticos grande concentração nos níveis mais altos e os saudáveis conecentrados em torno de 100.

Análise de correlação

A correlação descreve como as mudanças em uma variável estão associadas às mudanças em outra variável. A correlação é expressa por um coeficiente de correlação, que varia de -1 a 1. Um coeficiente de correlação próximo de 1 indica uma forte correlação positiva, o que significa que as duas variáveis tendem a aumentar juntas. Um coeficiente de correlação próximo de -1 indica uma forte correlação negativa, onde uma variável tende a diminuir quando a outra aumenta. Um coeficiente de correlação próximo de 0 indica que não há correlação linear entre as variáveis.

A correlação é importante na análise de regressão porque ajuda a entender a relação entre as variáveis independentes e a variável dependente. Antes de construir um modelo de regressão, é crucial examinar a correlação entre as variáveis independentes e a variável dependente. Se houver uma correlação forte entre uma variável independente e a variável dependente, isso sugere que a variável independente pode ser um bom preditor da variável dependente e pode ser incluída no modelo de regressão.

A seguir é possível visualizar a correlação entre as variáveis que estamos trabalhando:



Pode-se obervar que, em geral, as correlações entre as variáveis são baixas e todas positivas, sendo as mais correlacionadas oa pares *outcome* e *glucose*, *age* e *pregnancies* e o restante com medidas entre 0 e 0.3. Como nenhuma correlação, é, em módulo, maior que 0.9, não precisamos remover nenhuma covariável.

Análise de Regressão

Modelo

Um modelo de regressão linear visa descrever a relação entre uma variável dependente (também chamada de variável de resposta) e uma ou mais variáveis independentes (também conhecidas como preditoras ou explicativas). Para selecionar o modelo que melhor se ajusta aos dados, vamos usar como base o Critério de Informação de Akaike (AIC), que é uma medida que apresenta menor valor para o melhor modelo. Logo, nosso objetivo é encontrar o modelo que apresenta o menor AIC.

Para isso, vamos utilizar o método de seleção de variáveis chamado *Stepwise*, que começa com o modelo completo, ou seja, com todas as variáveis no modelo e remove ou adiciona as variáveis caso uma dessas opções gere uma diminuição no AIC. Quando nenhuma dessas ações ocasiona um menor AIC, consideramos que o melhor modelo foi obtido.

Em R, executamos esse processo primeiro utilizando a função `lm()`, que ajusta o modelo aos dados, estimando os coeficientes que melhor ajustam os dados observados. Em seguida, usamos a função, `stepAIC()`, com argumento `direction="both"`, que funciona seguindo um procedimento iterativo que envolve adicionar ou remover variáveis independentes do modelo, uma de cada vez, e comparar os valores do AIC para determinar se a adição ou remoção da variável resulta em uma melhoria no ajuste do modelo. O processo continua até que nenhuma alteração adicional resulte em uma redução significativa no AIC.

Ao executar esse processo aplicado aos dados do Instituto Nacional de Diabetes e Doenças Digestivas e Renais, obtemos que as variáveis presentes no modelo para descrever a variável *glucose* são *pregnancies*, *bllod_pressure*, *bmi*, *age* e *outcome*, com coeficientes -0.69186, 0.23415, 0.29128, 0.44003 e 27.73519, respectivamente.

Verificando pressupostos do modelo

Os pressupostos de um modelo de regressão são condições que devem ser atendidas para garantir a validade e a confiabilidade das inferências e previsões feitas pelo modelo. Esses pressupostos são fundamentais para assegurar que as estimativas dos coeficientes do modelo sejam precisas e que os testes estatísticos sejam válidos. Aqui vamos verificar os pontos influentes e de alavanca, a normalidade dos erros e a homocedasticidade. Garantir que esses pressupostos sejam atendidos é essencial para a construção de modelos de regressão robustos e confiáveis, que possam ser usados para tomar decisões informadas e baseadas em dados.

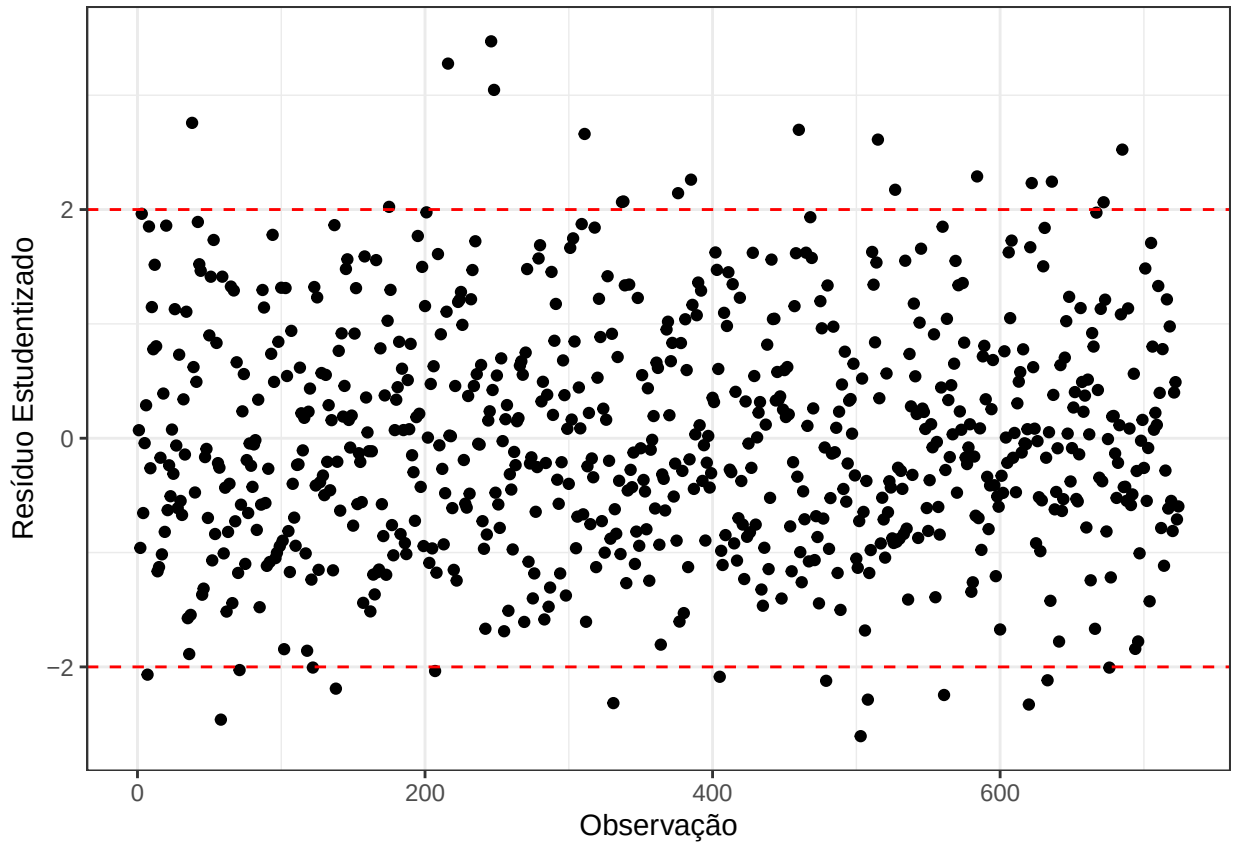
Multicolinearidade

O pressuposto da multicolinearidade refere-se à situação em que duas ou mais variáveis independentes em um modelo de regressão estão altamente correlacionadas entre si. Em outras palavras, existe uma relação linear quase perfeita entre algumas das variáveis independentes. Esse fenômeno pode ser prejudicial para a interpretação do modelo e para a precisão das estimativas dos coeficientes de regressão. Quando existe multicolinearidade, torna-se difícil interpretar os efeitos individuais de cada variável independente sobre a variável dependente.

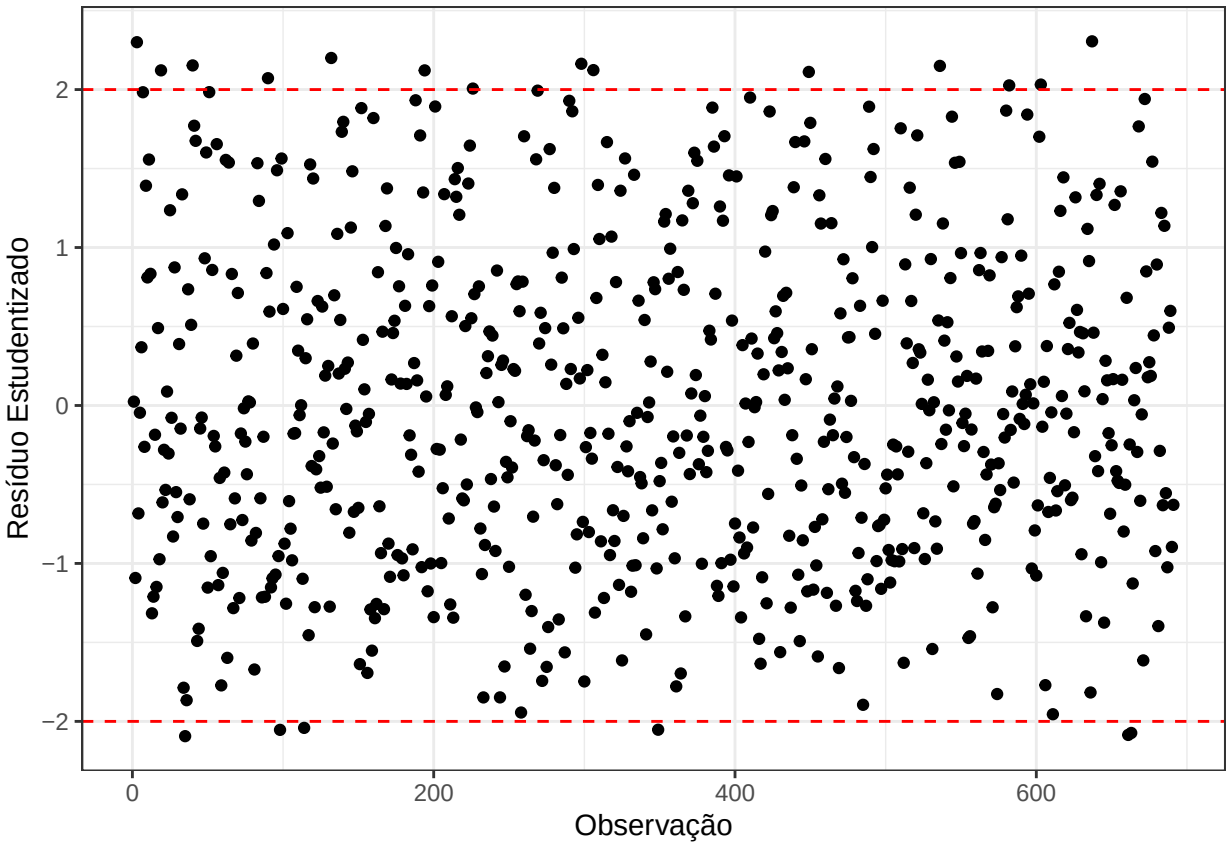
Já vimos pelo gráfico de correlação, que não há nenhum par de variáveis altamente correlacionadas, agora vamos verificar, usando a VIF, uma métrica que quantifica o grau de multicolinearidade entre as variáveis independentes. Valores de VIF maiores que 10 são frequentemente considerados indicativos de multicolinearidade significativa. Usando a função `vif()`, do R, observamos que os valores variam de 1.187740 a 1.591839, logo não há multicolinearidade.

```
##    pregnancies blood_pressure            bmi            age        outcome
##       1.471071       1.227783       1.199904       1.591839       1.187740
```

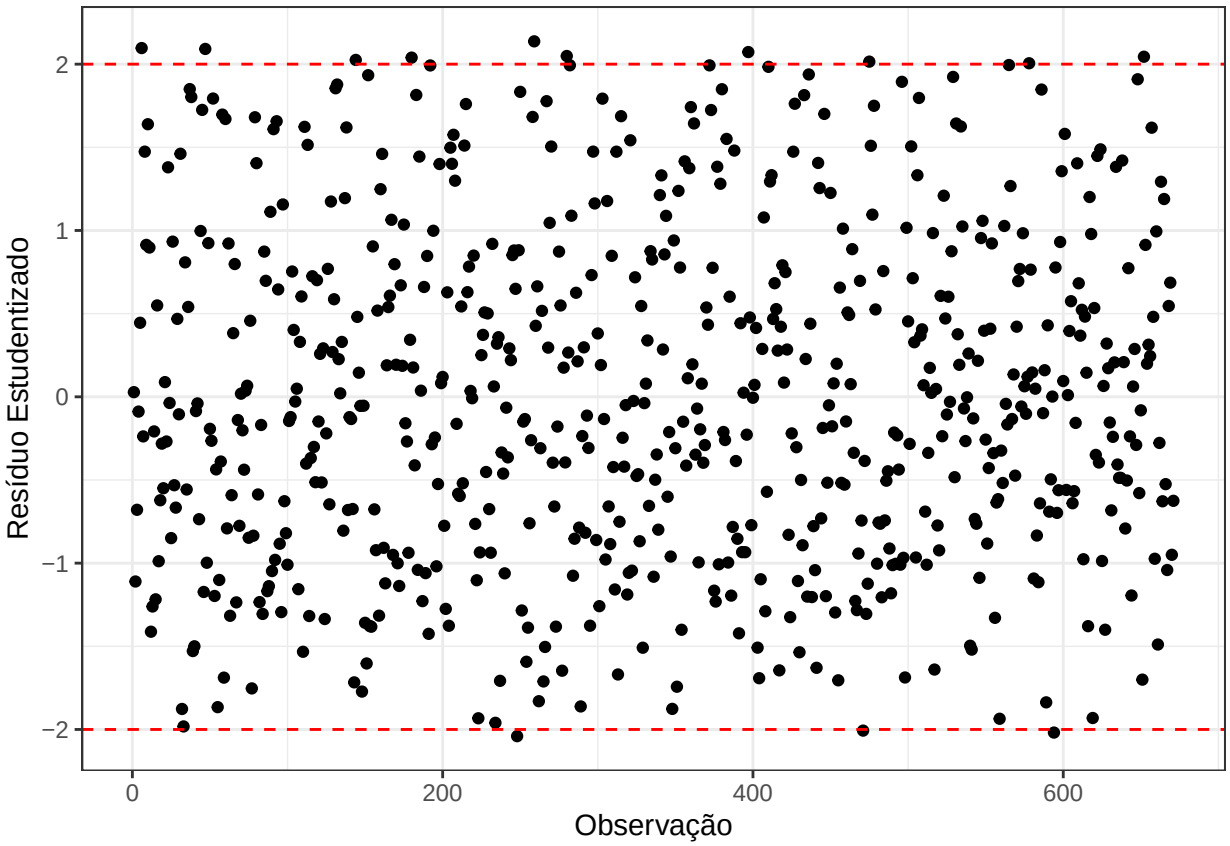
Identificação de Outliers Valores discrepantes presentes nos dados podem interferir na correta estimação do modelo, assim, é necessário removê-los. Para identificar esses pontos atípicos, utilizamos os resíduos estudentizados, que são uma medida padronizada dos resíduos de um modelo estatístico calculada dividindo o resíduo pelo seu desvio padrão estimado. Essa padronização permite que os resíduos sejam comparados entre si e com valores de referência, facilitando a identificação de observações incomuns ou atípicas no conjunto de dados. Aqui vamos definir os "limites" para esses resíduos estudentizados como 2 e -2, retirando os dados com resíduos fora do intervalo $[-2, 2]$. A seguir representamos esses resíduos graficamente.



Aqui identificamos as observações discrepantes e vamos retirá-las, reconstruindo o modelo e agora obtendo o seguinte gráfico de resíduos estudentizados:



É possível notar que ainda há a presença de alguns *outliers*, assim, vamos executar o processo de remover as observações e construir o modelo novamente, obtendo:

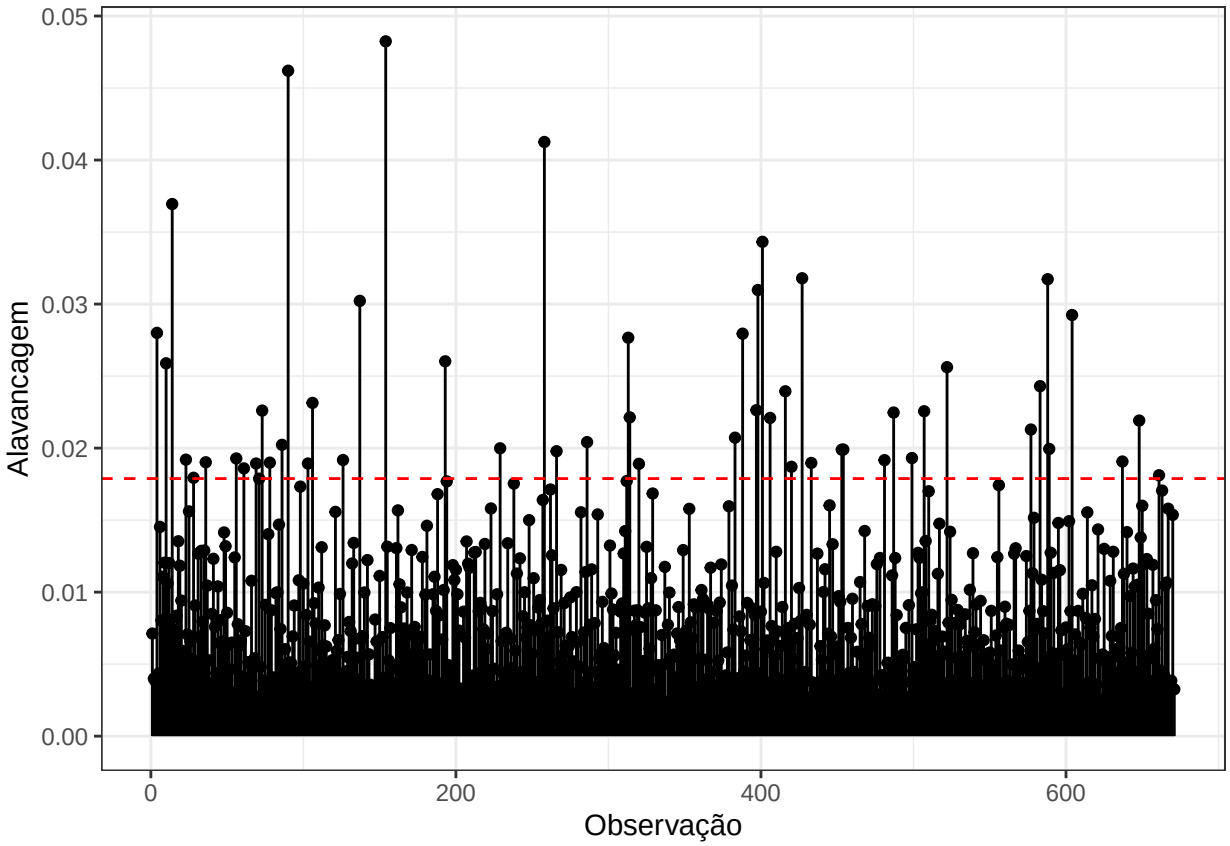


Acima podemos notar que há poucos outliers e que esses estão localizados bem próximos aos "limites" que adotamos. Assim, podemos seguir verificando os demais pressupostos do modelo.

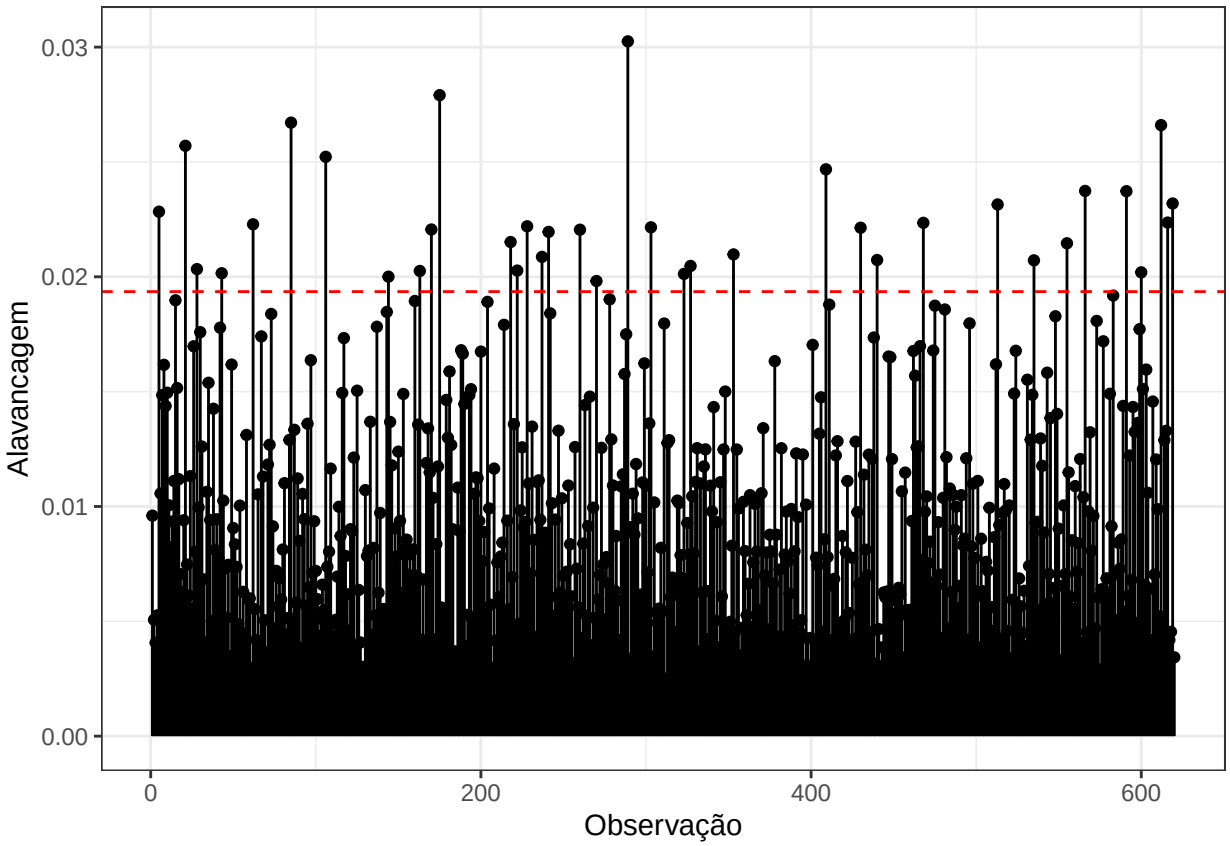
Pontos de alavanca

Pontos de alavanca são observações que possuem valores extremos nas variáveis independentes (preditoras). Esses pontos têm o potencial de influenciar significativamente a posição da linha de regressão, devido à sua distância em relação à média das variáveis independentes.

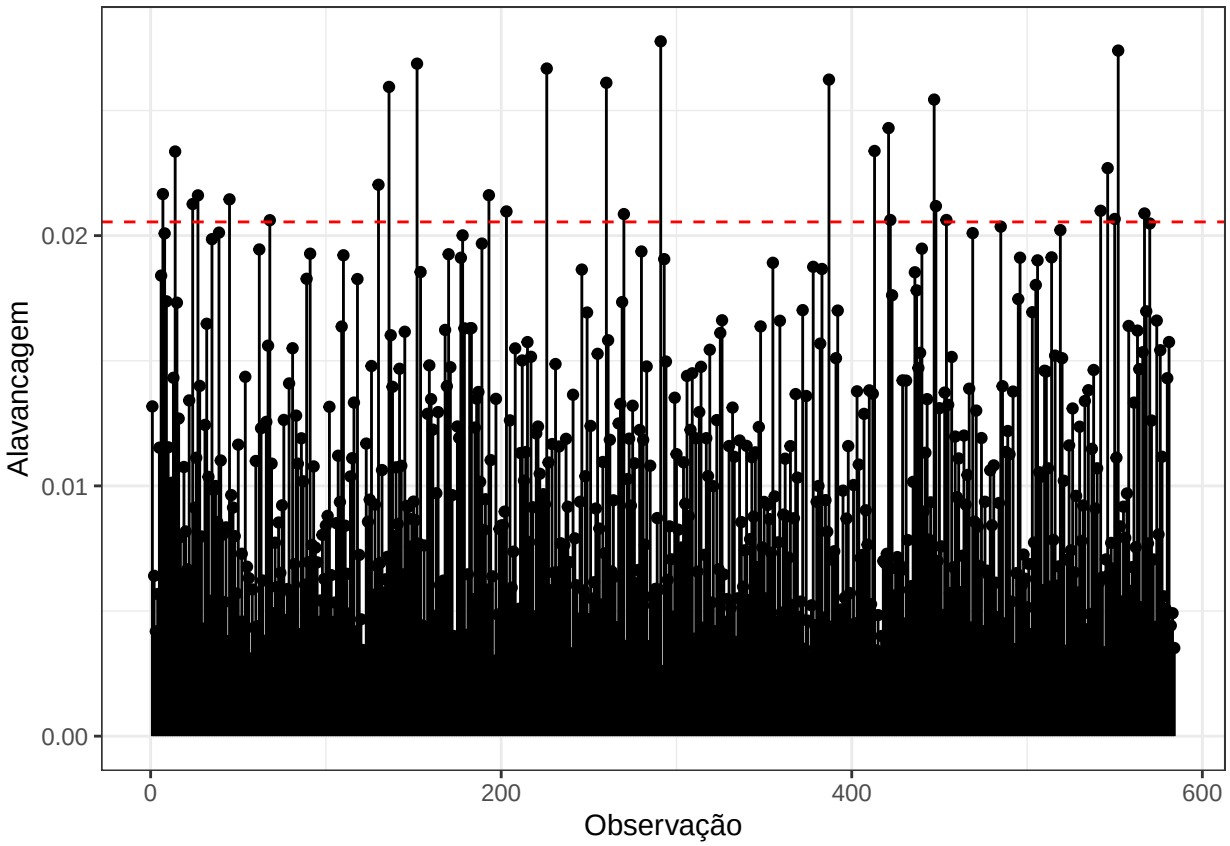
Para identificar os pontos de alavanca vamos analisar se o valor de alavanca (calculado a partir da matriz H) de cada observação excede $\frac{2k+1}{n}$, em que k é o número de covariáveis e n o número de observações.



Aqui percebemos que muitas observações ultrapassam o valor limite estipulado, assim removemos esses dados e reconstruímos o modelo, obtendo:



Como ainda há valores que ultrpassam o valor de corte, vamos executar esse processo novamente:

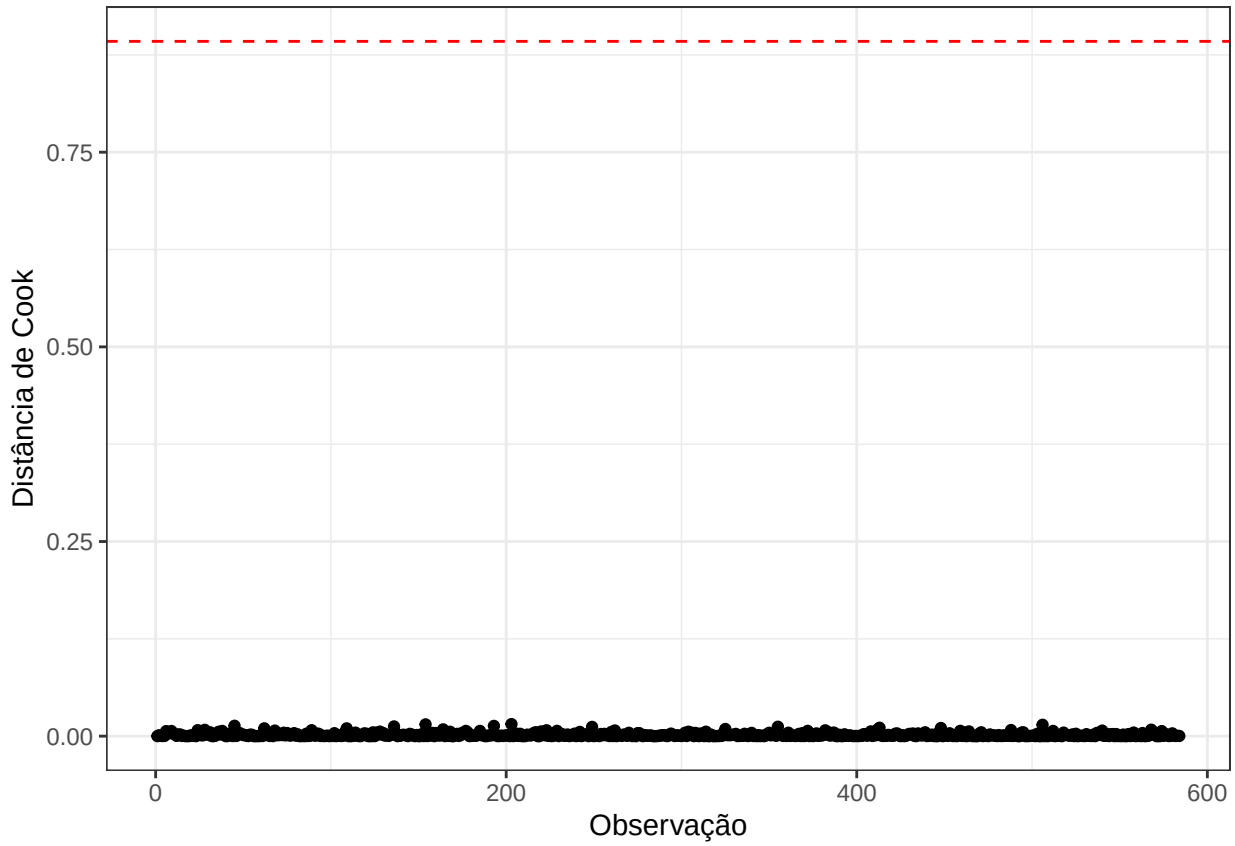


Aqui há menos observações acima do valor de corte e que ultrapassam esse limite em menor valor. Assim, vamos seguir para os demais pressupostos.

Pontos influentes

Pontos influentes são observações que têm um impacto significativo nos parâmetros estimados do modelo, podendo distorcer os resultados da regressão e levar a conclusões incorretas sobre a relação entre as variáveis. Identificar e avaliar a influência desses pontos ajuda a garantir a robustez do modelo e a confiabilidade das inferências feitas a partir dele.

Uma métrica comum para identificar pontos influentes é a distância de Cook, que combina tanto a magnitude do resíduo quanto a alavanca de uma observação para determinar sua influência no ajuste do modelo. Valores próximos de 0 indicam que a observação não é influente, maiores que 1 sugerem que a observação pode ser influente e entre 0 e 1 indicam que a observação pode ter alguma influência, mas não é necessariamente preocupante. A seguir estão essas distâncias representadas graficamente:



Como nenhuma observação ultrapassa o ponto de corte estipulado, não precisamos fazer nenhuma alteração no modelo vigente.

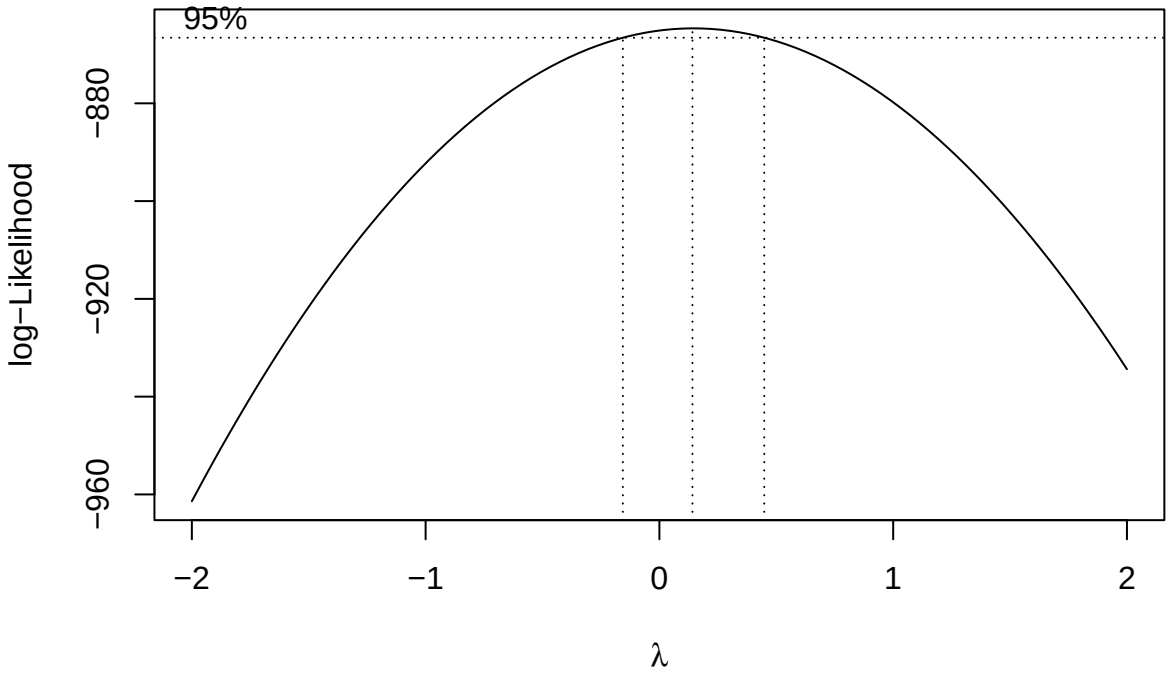
Normalidade

O pressuposto da normalidade requer que os erros do modelo sejam distribuídos de forma aproximadamente normal. Este pressuposto é crucial, especialmente quando se deseja realizar testes de hipóteses e construir intervalos de confiança para os coeficientes de regressão.

Um dos procedimentos para avaliar se os residuos são normalmente distribuídos é o teste Lilliefors (Kolmogorov-Smirnov), que é usado para verificar se uma amostra de dados segue uma distribuição normal.

```
##                                             Estatística      p
## Lilliefors (Kolmogorov-Smirnov) normality test       0.039 0.0346
```

Considerando um nível de significância de 0.05, rejeitamos a hipótese de normalidade dos resíduos, já que o valor-p é 0.0346, menor que 0.05. Para cumprir o pressuposto e obter resíduos normalmente distribuídos, vamos usar a transformação de Box-Cox, que é uma técnica utilizada para transformar dados não normais em uma forma que siga uma distribuição aproximadamente normal, sendo útil na modelagem estatística e na análise de regressão para garantir a normalidade dos resíduos. Vamos escolher o valor de λ para a transformação de forma que maximize a função de verossimilhança, mostrada graficamente abaixo:



Assim, aplicamos a transformação para um $\lambda = 0.1414141$ e realizamos o teste novamente, obtendo:

```
##                                            Estatística      p
## Lilliefors (Kolmogorov-Smirnov) normality test      0.0355 0.0766
```

Aqui obtemos um p-valor maior que 0.05, assim aceitamos a hipótese de que os resíduos são normalmente distribuídos.

Homoscedasticidade

A homocedasticidade ocorre quando a variância dos erros (ou resíduos) é constante ao longo de todos os níveis das variáveis independentes. Quando esse pressuposto é satisfeito, dizemos que os erros são homocedásticos. Caso contrário, quando a variância dos erros varia, temos heterocedasticidade. A presença de heterocedasticidade pode levar a estimativas de coeficientes com variâncias subestimadas ou superestimadas, resultando em testes de significância e intervalos de confiança incorretos.

Para verificar a homocedasticidade, uma das ferramentas utilizadas é o Teste de Breusch-Pagan, que avalia a hipótese nula de que os erros têm variância constante. Um valor de p pequeno (tipicamente menor que 0,05) indica a presença de heterocedasticidade. Realizando esse teste em R obtemos o resultado:

```r
bptest(fit6)                    # teste de Breusch-Pagan
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  fit6
## BP = 7.2478, df = 5, p-value = 0.2029
```

Assim, aceitamos a hipótese nula de homocedasticidade.

Erros não-correlacionados

O pressuposto de erros não-correlacionados significa que os resíduos (ou erros) do modelo devem ser independentes uns dos outros. Ou seja, o erro associado a uma observação não deve fornecer qualquer informação sobre o erro associado a outra observação.

Para verificar a independência dos erros em um modelo de regressão, vamos usar o teste de Breusch-Godfrey, que examina a hipótese nula de que não há autocorrelação nos resíduos. A hipótese alternativa é que existe autocorrelação entre eles.

```
bgtest(fit6)                    # teste de Breusch-Godfrey
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  fit6
## LM test = 0.83245, df = 1, p-value = 0.3616
```

Nesse caso, como obtemos um p-valor maior que o nível de significância 0.05, aceitamos a hipótese de erros não correlacionados.

Conclusão

Neste trabalho, abordamos a construção de um modelo de regressão para a variável glucose em um conjunto de dados coletados pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais. O estudo envolveu diversas variáveis preditoras como *pregnancies*, *bloodPressure*, *skinThickness*, *insulin*, *bmi*, *diabetesPedigreeFunction*, *age* e *outcome*. Através dessa análise, destacamos a importância de verificar e satisfazer os pressupostos fundamentais da regressão linear para garantir a validade e a precisão das inferências estatísticas.